Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2017-273-RC3, 2017
© Author(s) 2017. This work is distributed under
the Creative Commons Attribution 4.0 License.

Geoscientific Model Development Discussions

Interactive comment on "On the Effect of Model Parameters on Forecast Objects" *by* Caren Marzban et al.

Anonymous Referee #3

Received and published: 20 December 2017

General comments

In this paper, the authors develop a framework for conducting sensitivity analysis (SA) for the output of numerical models, when the output of interest is a spatial field or realization of a non-smooth or non-continuous variable. In this case, the authors propose conducting the SA on features of "objects," which can be specified generally but here correspond to high quantile clusters of grid cell values of daily precipitation. Two statistical methods are used to determine clusters, namely Gaussian Mixture Model (GMM) clustering and Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and the authors explore sensitivity of their results to the clustering method. Features of these clusters are extracted and fed into a Multivariate Multiple Regression (MMR) model to estimate the effect of individual model parameters on the cluster features.

This is an interesting paper, and describes methodology for an important problem in sensitivity analysis, namely conducting SA for a spatial field of model responses. Furthermore, the clustering both addresses the nature of precipitation data (i.e., non-smooth or non-continuous data) while also reducing the dimension of the problem (i.e., considering fixed clusters rather than a fine spatial grid). The paper is well-written and nicely motivates the work, however, additional detail should be given in the section describing the statistical model, and I think reorganization of Section 3 would greatly improve the presentation (see the Technical corrections below). Furthermore, I am concerned with the analysis methods, particularly the significance testing, and am worried that the way in which the results are presented might be misleading (see the Scientific comments for more details).

Scientific comments

As a statistician, I will primarily comment on the statistical model and significance testing, leaving discussion on the experimental design of the sensitivity analysis and variables selected for analysis (i.e., latitude, longitude, intensity, area, orientation, and eccentricity) to more informed parties. In my opinion, the clustering approaches considered (GMM and DBSCAN) seem reasonable, and it was nice to see that results are robust to the clustering method used.

My first concern has to do with the description of the MMR model as well as the treatment of daily replicates within this model. The authors present a generic description of a multiple linear regression model in Equation (1), but it would be helpful to more clearly describe the generalization to the **multivariate** multiple linear regression model that was actually used. If I am following everything correctly, the statistical model you actually use is

$$\begin{bmatrix} y_t^{min} \\ y_t^{med} \\ y_t^{max} \end{bmatrix} = \begin{bmatrix} \alpha^{min} \\ \alpha^{med} \\ \alpha^{max} \end{bmatrix} + \begin{bmatrix} \beta_1^{min} \\ \beta_1^{med} \\ \beta_1^{max} \end{bmatrix} x_{1,t} + \cdots + \begin{bmatrix} \beta_{11}^{min} \\ \beta_{11}^{med} \\ \beta_{11}^{max} \end{bmatrix} x_{11,t} + \begin{bmatrix} \delta_t^{min} \\ \delta_t^{med} \\ \delta_t^{max} \end{bmatrix}$$

(where min = minimum, med = median, and max = maximum), or, written in vector form,

$$y_t = \alpha + \beta_1 x_{1,t} + \cdots + \beta_{11} x_{11,t} + \delta_t, \tag{1}$$

for $t = 1, \ldots, 99$ samples taken from the 11-dimensional parameter space. Presumably, you use the usual MMR assumption that the error vectors δ_t are independent and identically distributed as Normal with mean vector 0 and non-diagonal covariance matrix Σ (i.e., the elements of δ_t are correlated). Is this a correct characterization of the model?

In practice, you actually estimate the 3×11 β coefficients from Equation (1) for each of six features and each of 40 days, presenting boxplots of the β coefficients aggregated over the 40 daily replicates for each of the $3 \times 11 \times 6$ combinations of feature summaries/input variables/features. (See below for a concern related to the boxplots.) This seems like an unnecessary complication to the analysis. As evidenced by your decision to keep only every third day (reducing your data from 120 days to 40 days) in order to remove temporal correlation, it seems to me that these 40 days could represent an ensemble of realizations for each of the 99 parameter settings. Thus, instead of fitting 40 separate MMR models for each of the 6 features, your model could instead be

$$y_{td} = \alpha + \beta_1 x_{1,t,d} + \cdots + \beta_{11} x_{11,t,d} + \delta_{td} \tag{2}$$

for $d = 1, \ldots, 40$ days (I assume that $x_{j,t,d} = x_{j,t}$ for $j = 1, \ldots, 11$, i.e., that the input parameter settings are the same for each day). In other words, instead of using the 40 daily replicates to estimate the distribution of each β coefficient, you could build this variation into the statistical model and directly estimate the variability of the coefficients, then calculating P-values or confidence intervals as required. This seems to be a more refined way to handle the daily replicates, especially since it seems that you are not concerned with how the β coefficients vary across the different days.

Secondly, I am concerned by the significance testing procedure and the presentation of results. First of all, your two-stage procedure for controlling Type I error seems ad hoc, particularly your qualitative approach to assessing individual significance in the

second stage. The omnibus test in the first stage is a good idea (although it would be helpful to have more details given on exactly what you have done – instead of simply providing citations), but you need to be careful about the multiple testing even after reducing the number of tests to $6 \times 40 = 240$. I appreciate that you have at least considered a Bonferroni adjustment, but you should think carefully about this choice: Bonferroni controls a family-wise error rate, implying that the collective conclusion of all tests is invalid if at least one Type I error is made. I don't think this is actually what you want – it seems to me that you simply want to control the number of Type I errors. As an alternative, you might consider the very simple procedure for controlling the rate of false discoveries (i.e., FDR) given in the classic paper by Benjamini and Hochberg (1995). Their simple procedure is remarkably powerful and could more appropriately address the multiple testing issue.

Regardless, after you have conducted the omnibus test, you proceed to present box plots of the coefficient estimates, aggregated across the daily replicates. I think that such an aggregation of the coefficient estimates provides you with a sampling distribution of the true coefficient estimate – please correct me if this is not the right way to think about this. In any case, the aggregated coefficient estimates are most certainly not a posterior distribution of the true coefficient, which is what you would get from a fully Bayesian analysis. In this case, it is misleading to represent a sampling distribution with a boxplot: if the boxplot is skewed to the right, this does not mean that the distribution of the true coefficient is skewed to the right. Instead, you should represent sampling distributions using a confidence interval, which could be plotted as a box (with no whiskers) or a solid bar. Additionally, simply checking to see if boxplots overlap with zero is not an appropriate way to assess *statistical* significance: what significance level is being considered?

My suggestion would be to fold the daily replicates into the MMR as suggested in Equation (2), and calculate P-values for each of the $11 \times 3 \times 6$ coefficients. Then, I would use the Benjamini and Hochberg procedure to identify statistically significant

coefficients at a particular level α. Instead of the boxplots in Figures 2 and 3, I would recommend using points or bars to indicate the magnitude of the coefficient estimate and shading or masking to indicate which estimates are statistically significant.

Technical corrections

On a more technical note, I found the organization of Section 3 to be very confusing. I would suggest moving Sections 3(d) and 3(e) to immediately follow Section 3(a). In this case you will have already described the clustering and the features of interest before discussing the statistical model and significance testing. I would also recommend moving lines 161-170 into Section 3(e).